8. New Perspectives in the Analysis of Abilities

John B. Carroll
University of North Carolina at Chapel Hill

8 New Perspectives in the Analysis of Abilities

John B. Carroll
University of North Carolina at Chapel Hill

INTRODUCTION

One can understandably be skeptical when a ''new perspective'' is offered on a topic that has been under scientific examination for a very long time. I am not sure that I have any truly new perspectives, but I entertain the notion that my perspectives have the kind of novelty that will last long enough to permit taking a fresh look at some very old problems and getting new insights into their solution. I'm concerned with several such problems: First, what is an ''ability''? How can an ability be defined? This is a problem that I believe has never been adequately addressed in the psychometric literature. Second, how can data from ability measurements be best analyzed to help in the definition of the ability, and thus to determine what has often been called the ''construct validity'' of the measurements? Third, what are the implications for the construction of better measurements of ability? Throughout my presentation, one detects influences from cognitive psychology—influences that I point out, but my primary concern is with psychometric aspects of ability measurements.

Here, I use the term ''ability'' in a very general sense, so that it covers both the concept of aptitude and the concept of achievement. At the stage of *defining* an ability, the difference between aptitude—thought of as a capacity for some future achievement—and achievement—thought of as the demonstration of some acquired performance—is irrelevant, because, as will shortly be seen, we are concerned in either case with deriving the definition of an ability from observations of performance. The question of the source of the performance (i.e., to what extent it comes through constitutional/genetic factors and to what extent it comes through learned experiences) need not enter discussion.

WHAT IS AN ABILITY?

The *American Heritage Dictionary* defines ability as "the quality of being able to do something; physical, mental, financial, or legal power to perform." We can immediately drop consideration of financial and legal powers; from the standpoint of psychological and educational measurement we can be concerned, however, with physical and mental powers. Nevertheless, even the definition offered by the dictionary has an air of circularity: Ability is defined in terms of "being able to perform something," and ironically enough, the word *able* is defined in terms of "having sufficient ability." I'm afraid the dictionary is of little help in defining "ability," except possibly in the phrase "ability to perform something." What is this *something?* In the context of psychological and educational measurement, must it not refer to some *class* of tasks? If we think of commonly recognized abilities such as athletic ability, or musical ability, the common assumption is that a person with such an ability is able to perform well a variety of tasks that can be called athletic, or musical, as the case may be. When psychologists and educators speak of "mental ability," they are referring to performance in a variety of "mental" tasks. The question is, what is a "mental" task?

We know that abilities are often of a more specialized character. A good basketball player is not necessarily a good 100-yard runner; a good pianist is not necessarily a good composer, or not even a composer at all. Evidence from factor-analytic investigations of mental abilities suggests that there exist a number of somewhat unrelated mental abilities: verbal ability, reasoning ability, spatial ability, numerical ability, and so on. Correlational and factor-analytic evidence is of some use in classifying and identifying abilities, because it yields information on what abilities are likely to go together or to be separate. More precisely, it yields information on the classification of the *tasks* that call for different abilities.

Let us focus on the fact that the tasks that call for a particular ability, whatever it is, can be of considerable variety, perhaps even of infinite variety. How do they vary? One dimension along which they vary is their *difficulty*. One can often diagnose what causes tasks to vary in difficulty. In simple cases, it may be a matter of physics or physiology. In basketball, it is harder to shoot a basket from a long distance than from a short distance. In musical performance, Bach Inventions are generally much easier to play than most of the compositions of, say, Debussy. In fact, it has long been the practice of music educators to assign grades of difficulty to instrumental musical compositions; I do not know whether anyone has analyzed exactly what makes for ease or difficulty of such compositions.

In psychological and educational measurement, the concept of difficulty turns up in the form of information about the proportions of tested samples or populations that are able to "pass" each of the items on a test. Such information is often used in arranging the items of a test in order of difficulty, apparently on the assumption that subjects will be more comfortable in taking the test if they can start with easy tasks. But as in the case of musical compositions, there is usually little concern with what makes items easy or difficult. Test makers often simply take item difficulty data as givens that need not be questioned further. Note, by the way, that ''items'' on a psychological or educational test are really ''tasks'' that call for correct performance; the more of these tasks the examinee can perform correctly, the higher the score, and the higher the level of ''ability'' that is inferred from the score.

A preliminary evaluation of the ''construct validity'' of a test is often made simply by considering the class of tasks that is involved in the test. If all the items are concerned with English spelling, for example, the test may be regarded as a test of ''spelling ability.'' Or if all the items seem to involve ''manipulation of visually presented spatial relationships,'' the test is regarded as a test of ''spatial ability.'' But intuitive classifications of tasks are often unsatisfactory, perhaps by their very nature. They yield no guarantee that there is only one spelling ability, or only one spatial ability. In the case of spatial ability, at least, the available evidence is to the contrary (Lohman, 1979).

At the same time, it is often pointed out that it is difficult to establish the unitary or nonunitary nature of abilities from correlational studies of items or tasks. The difficulties are technical, stemming from problems with the interpretation of bivariate distributions of item responses. Much of our knowledge about the differentiation of abilities comes from factor-analytic studies using scores on multi-item tests, the tests being composed of series of plausibly similar items. There is now some promise in recently developed techniques for item factor-analysis (Wilson, Wood, & Gibbons, 1984) but as yet these techniques have not been widely applied, and I myself have not yet had the opportunity to use them.

But I am getting off the track. Suppose, for the sake of argument, that we have a set of tasks that can be demonstrated to measure a single ability at different difficulty levels. What might convince us that they measure a single ability would be evidence that there are systematic relationships between characteristics of individuals and the levels of difficulties of the tasks, such that individuals who *can* perform the more difficult tasks have a uniformly higher probability of passing the easier tasks than those who can *not* perform the more difficult tasks, and also, such that individuals who cannot perform the easy tasks also cannot perform the harder tasks. This idea is not new; to my knowledge it was first pointed out by David Walker, a Scottish educational psychologist, in a series of papers published in the *British Journal of Psychology* over the years 1931 to 1940 (Walker, 1931, 1936, 1940). Walker called tests having the above-mentioned property ''unig,'' whereas tests not having this property were called ''hig'' (from the expression ''higgledy-piggledy''). Walker anticipated the idea of what later came to be known as the Guttman scale, and I like to refer to it as the Walker scale, or perhaps the Walker-Guttman scale.

There is much more to this idea, however, I can best illustrate it by referring to data that I collected some years ago on a test that I believe can be shown to measure a single dimension of ability, namely musical pitch discrimination ability. This is the old Seashore Sense of Pitch test; in fact, the data I collected were for the 1919 version of the test. Let me describe this test, in case you are not familiar with it. It consists of 100 items, divided into ten subsets of 10 items each. Each item in a given subset presents, by a phonograph recording, two tones that differ in pitch by a certain amount, constant over the items in the subset; the subject's task is to indicate on the answer sheet whether the second tone is higher or lower than the first. The pitch difference in the easiest subset is 30 cycles per second, or (considering the overall pitch level) about a semitone; the pitch differences in other subsets range down to one-half cycle. Subjects are required to make a response to each item, and thus there is obviously an element of guessing, or success by chance, of 50%. I may note, incidentally, that some years ago Guilford (1941) (while he was in the psychology department at the University of Nebraska) collected and analyzed data with this test and claimed that the test measures three separate abilities. I have recently shown, however (Carroll, 1983) that Guilford was misled by statistical artifacts, and that the test measures essentially only one ability. Imperfections in the 1919 recording add a certain element of response set bias, but this may be ignored for practical purposes.

In a further analysis of the data I collected on about 1100 college students, I wanted to study curves of performance in relation to the pitch differences of the subtests. How did the curves of performance for students with high scores compare with those for students with average and low scores on the test? I divided the total score distribution into deciles and plotted average performance curves for each decile. The results are shown in Fig. 8.1. The baseline is scaled in terms of the logarithm of the pitch difference; the ordinate shows the probability of correct performance. As may be seen, the data are quite systematic. High ability students have practically perfect performance for subtests with large pitch differences; their average performance descends to a threshold only at a pitch difference of about 1.25 Hertz, the limen or threshold being set at 75% correct (halfway between perfect and chance performance). Students in the lowest decile of ability, on the other hand, have on the average a threshold performance at a pitch difference of about 20 Hertz.

The curves have, as one might expect, the general shape of normal ogives, and have approximately the same slope. This slope can be expressed in terms of the logarithm of the pitch difference: one standard deviation of the response curve is about .25 log pitch difference units. I believe that this slope is in fact *characteristic* of pitch discrimination ability. Even with a better-recorded test, and with many more items, this slope would probably not change much. The fact that the slope is not higher, as it would be if the slopes were as represented in Fig. 8.2, puts a certain constraint on the reliability of any test of pitch discrimina-

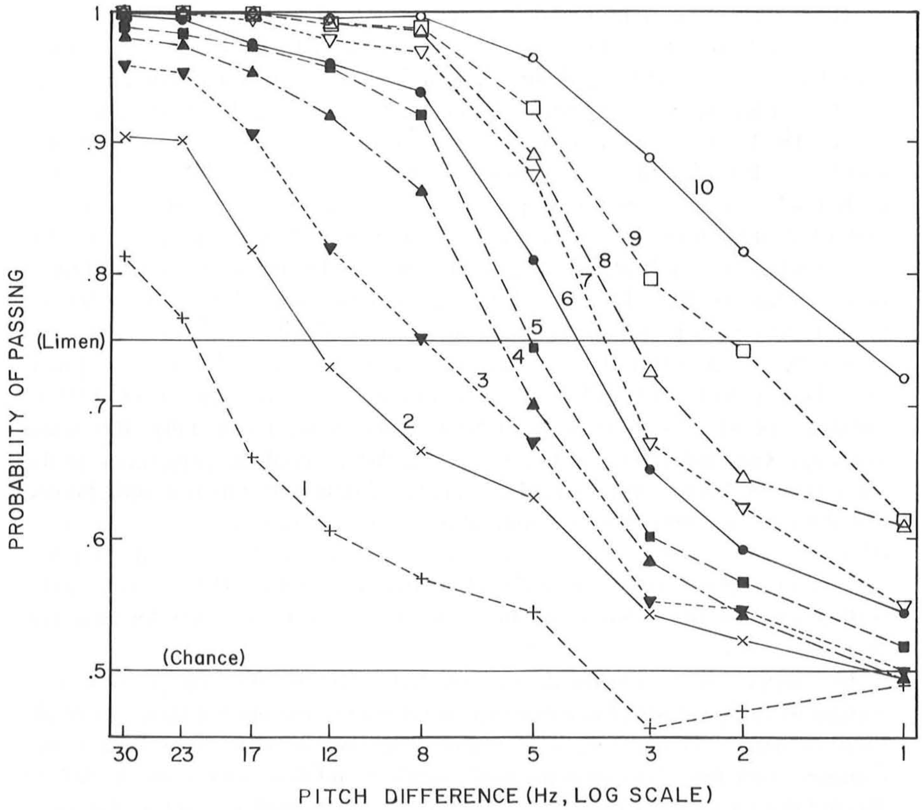

FIG. 8.1. Person characteristic functions for deciles of the total score distribution, Seashore Sense of Pitch test (N = 1082).

tion ability. At least, it puts a constraint on the reliability-per-item, and thus on the reliability of a test of any given length. (It is possible, in fact, to specify the limits on reliability in terms of parameters of the slope function.)

Of even more importance is the fact that these data support the *existence* and *definition* of pitch discrimination ability, in the sense that pitch discrimination ability is revealed in a systematic relation between individual characteristics and performance on subtests of different pitch difference levels. What makes for "difficulty" in pitch discrimination is the smallness of the pitch difference. High ability individuals have much smaller pitch difference thresholds than low ability individuals.

These data illustrate a paradigm that I believe can be transferred or applied to *any* ability. That is, an ability—any ability—can be defined in terms of the relation between individual thresholds of performance and the characteristics of tasks of different degrees of "difficulty." In the case of pitch discrimination

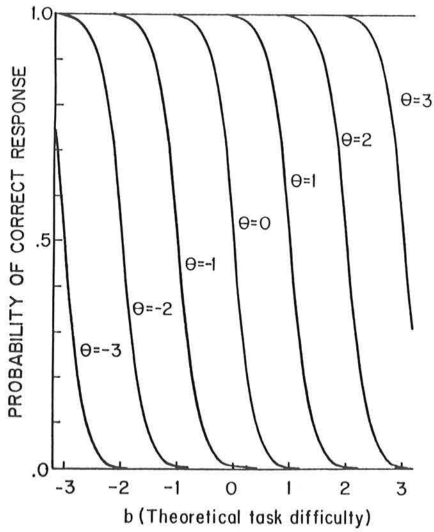

FIG. 8.2. Theoretical person characteristic functions for an ability with a high value of the slope parameter (a = 3).

ability, it is clear that the task characteristics are described in terms of pitch differences, and individual differences can be referred to threshold points on the pitch difference scale. What about other abilities?

To introduce this topic further, I present one other set of data, this time on a Block Counting test that was administered by my colleagues at the University of North Carolina (Johnson & Meade, personal communication) to 10th-grade-children in a study of the development of spatial abilities. The Block Counting test has been regarded as a test of some kind of spatial ability. The test used in this study is a little different from other block-counting tests that appear in some test batteries. Sample items are shown in Fig. 8.3. Each item is a perspective drawing of a pile of blocks; the subject's task is simply to count the blocks and write down the answer. Subjects are told that all blocks in a given drawing are of the same shape. Because the answers are free responses, there is practically no guessing element.

In Fig. 8.4 are shown average probabilities of correct answers for sets of items of varying difficulties, for *ninths* (noniles) of the total score distribution for 119 10th graders. Again, the data are quite systematic. High scoring individuals get correct answers on most of the "easy" items, and have only a little trouble with the hard items. Low scoring individuals have trouble even with the easy items, and have very little chance of passing the hard items. One can specify thresholds of performance for different individuals. Beyond stating it in terms of difficulty level, however, the baseline scale cannot easily be described. We must study, therefore, what makes the items easy or hard, since whatever makes for task difficulty is what gives rise to differences in ability, and thus leads toward a definition of that ability.

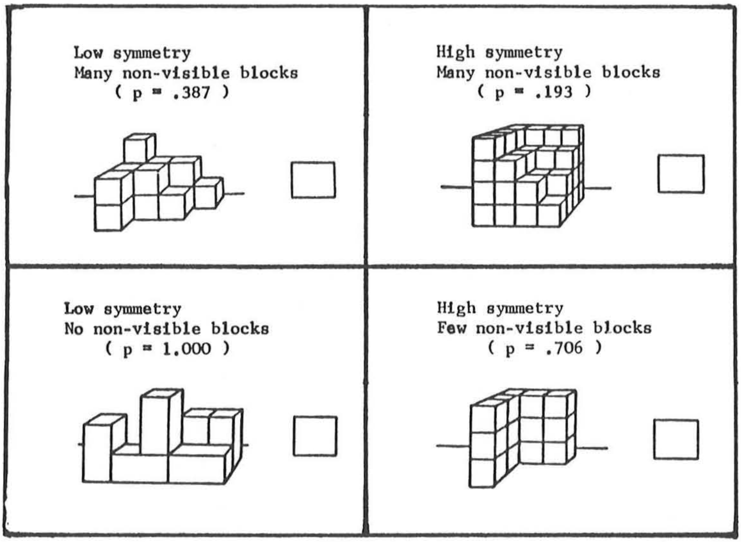

FIG. 8.3. Sample items from a Block Counting test, arranged to suggest the effect of "symmetry" and of the proportion of nonvisible blocks on item difficulty (p = proportion of 10th grade students giving correct answer). Copyright 1986 by Industrial Psychology Inc., 515 Madison Avenue, New York, NY 10022. All rights reserved. Permission granted for limited reproduction in professional psychometric journal in this instance only.

Detailed examination of the items, arranged in order of difficulty (proportion failing), discloses that they vary mainly in two characteristics: (1) the proportion of blocks that are not "visible" because they are hidden by other blocks, and (2) what I call the "symmetry" of the piles, that is, a characteristic such that one can use arithmetic computations to arrive more quickly at the number of blocks. The first of these variables has the greatest influence on item difficulty, but it interacts with the second. In Fig. 8.3 I have arranged the 4 sample items in such a way as to suggest how these task characteristics affect item difficulty. The two items in the bottom row have no or few nonvisible blocks, and are relatively easy, while those in the top row have many nonvisible blocks and are harder. The items in the left column have little "symmetry"; the subject must simply count the blocks more or less one by one. The items in the right column have high symmetry, and counting the blocks can involve some simple arithmetic. For example, the item at the lower right appears to be composed of a wall of $3 \times 3 = 9$ blocks at the left, plus an adjoining wall of $2 \times 3 = 6$ blocks, or 15 blocks in all. The items in the left-hand column are somewhat easier than those in the right-hand column.

From this analysis, it appears that the ability chiefly measured by this test is the ability to visualize the positions of blocks that are not immediately visible.

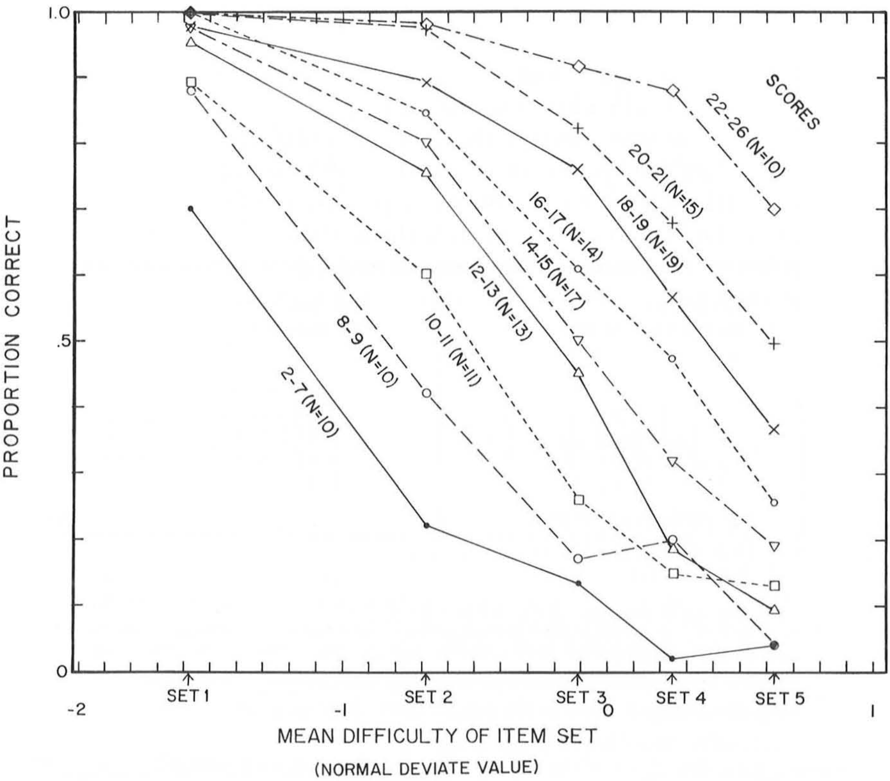

FIG. 8.4. Person characteristic functions for noniles of the total score distribution for the Block Counting test as given to 119 10th-grade children.

Secondarily, it measures an ability to use simple arithmetical processes in arriving at answers. As a matter of fact, the "symmetry" dimension in this task may tend to distort the assessment of the subject's ability to visualize missing blocks. Possibly a better, purer test of visualization ability could be devised by constructing all items with a minimal amount of symmetry, so as to reduce the possibility of using arithmetical processes.

Suggested by these findings, further questions could arise and be answered by appropriate investigations. Is the ability to visualize hidden blocks specific to the block counting task, or would it be found to be correlated with abilities in other types of visualization tasks, for example, the "surface development" test used by Thurstone (1938) or the mental paper folding test studied by Shepard and Feng (1972)? The answers to such a question could be found by analyzing data for the surface development and mental paper folding tests in the manner I have described, and examining relationships among the task parameters of the several tests.

THE PERSON CHARACTERISTIC FUNCTION (PCF)

The curves shown in Figs. 8.1 and 8.4—curves relating average performance of individuals to item difficulty—may be called *person characteristic functions* (PCFs). They have approximately the shape of normal ogives with a negative slope, descending from perfect or near perfect performance for "easy" items, through a threshold point, to zero or chance performance for "difficult" items. These curves are the reverse of the item characteristic curves familiar in item response theory. As a matter of fact, one can model these curves using precisely the same mathematical formula that is used in item response theory as developed by Lord (1980) and others. This is the three-parameter logistic function expressed as follows:

$$p = c + \frac{1 - c}{1 + \exp\left[-1.7a(\theta - b)\right]},$$

where p = the probability that an individual with ability θ will correctly perform an item or task characterized by the parameters a, b, and c, where

a = a parameter for the slope of the function;
b = a parameter specifying the difficulty of the item or task; and
c = a parameter specifying the probability that an individual completely lacking in ability ($\theta = -\infty$) will nevertheless perform the item or task correctly, as (often) by guessing.

The difference is that the person characteristic function plots performance *for an individual* (or group of individuals) as a function of item difficulty (the b parameter), whereas the item characteristic function plots performance *for an item* as a function of individual ability (the *theta* parameter θ). Both functions assume that all items measure the same latent ability (or cluster of abilities). Item characteristic functions have well-known uses in test theory, as Lord (1980) has shown. Use of the person characteristic function was first explored by Mosier (1941), although he did not call it that. The advantage I see for it is that it emphasizes the relation between ability and item or task difficulty. When there is a definite relation between ability and item difficulty, one is encouraged to explain that relation in terms of the characteristics of items or tasks.

Item response curves can also be used to look at these relations, but in this case one has to compare the functions for different items. This may be illustrated with data that I developed for vocabulary (opposites) items in the SAT, as shown in Fig. 8.5. (The data available to me did not permit computing person characteristic functions.) What we see in Fig. 8.5 are item characteristic curves for 15 vocabulary items; performance (in terms of percentage correct) is plotted against

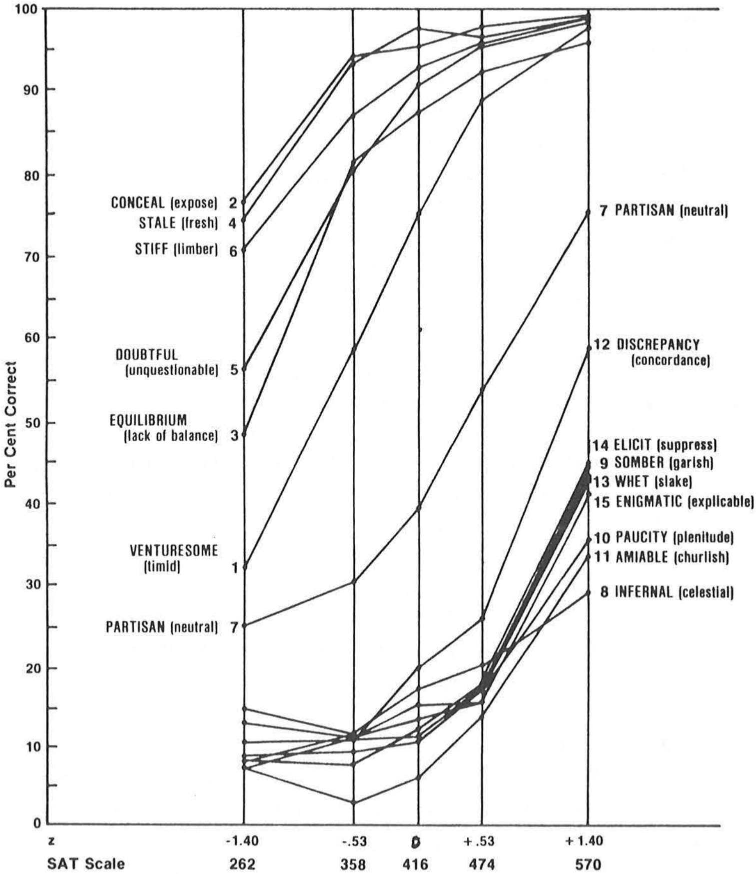

FIG. 8.5. Item characteristic curves for 15 verbal opposites items from a form of the SAT-V (from Carroll, 1980).

5 ability levels, actually quintiles (fifths) of an item analysis sample of 1920 cases with a mean SAT score of 416 and a standard deviation of 110. Obviously, as ability increases, correctness of performance increases; the curves are generally of a normal ogive shape with a positive slope. Note, however, that for most of the more difficult items, the percentages correct for low-scoring groups are well below chance levels (chance being 20% since these are 5-alternative items). The curves tend to have a U-shaped concavity, possibly because it is the low-average group, at an average SAT-V score of 358, that is most likely to be

seduced into choosing an incorrect alternative. The very low scorers don't even have enough ability to be seduced in this way; they are the ones who are most likely to answer by guessing.

The difficulty of the items can be measured in either of two ways: by the "delta" value derived from overall percentage correct, or by the threshold value estimated from the item characteristic curve. These two measurements are highly correlated, though not perfectly. What are the task characteristics of the items that make for difficulty? I estimated the familiarity of the words in the "lead" and the correct choices by using "SFI" (standard frequency index) values from the *American Heritage Word Frequency Book* (Carroll, Davies, & Richman, 1971). From these indices, item difficulty as measured by ETS's "delta" could be predicted with a multiple R of .80 ($p < .01$). This finding supports, at least, the rather obvious conclusion that these items are measures of vocabulary knowledge. What is more, however, the analysis using word frequency statistics makes it possible to specify in rather exact quantitative terms the range of vocabulary knowledge exhibited by examinees of given levels of ability. For example, consider the vocabulary knowledge shown for the top fifth of the sample, with a mean SAT-V or 570. These people have no trouble with words like CONCEAL, STALE, STIFF, DOUBTFUL, EQUILIBRIUM, and VENTURESOME, and the keyed correct answers *expose, fresh, limber, unquestionable, lack of balance,* and *timid,* respectively. But I find it rather disturbing that they tend to have trouble with words like PARTISAN, DISCREPANCY, ELICIT, SOMBER, WHET, ENIGMATIC, PAUCITY, AMIABLE, and INFERNAL.

One other example from my analysis of SAT items is instructive. (These data are more fully discussed in Carroll, 1980.) Figure 8.6 shows item characteristic curves for 10 "verbal analogies" items of an SAT-Verbal test. The common supposition is that these items measure "reasoning," that is, ability to discern an analogy. Sternberg (1977) developed a rather elaborate model for the behavior of solving analogies, involving among other processes the "encoding" of the stimuli, the "inference" of relations, and the "mapping" and the "application" of those relations. The question may be raised: To what extent do these processes make for difficulty of these items?

There is little evidence here that the examinees have difficulty with the concept and structure of an analogy per se. Even very low-scoring individuals have a fairly good chance of passing a simple analogy like number 27. This suggests that the SAT verbal analogies test does not measure the ability to solve analogies, as such, in the sense that low-scoring individuals would be less able than high-scoring individuals to deal with analogical structures, apart from their content. Instead, the evidence suggests that the harder items involve more difficult encodings, and more difficult and subtle inferences, mappings, and applications than the easy items. To a certain extent, there are vocabulary difficulties; Thus, low-scoring individuals probably have difficulty in encoding concepts represented by words like "slink," "furtive," and "innocuous." But the major difficulty

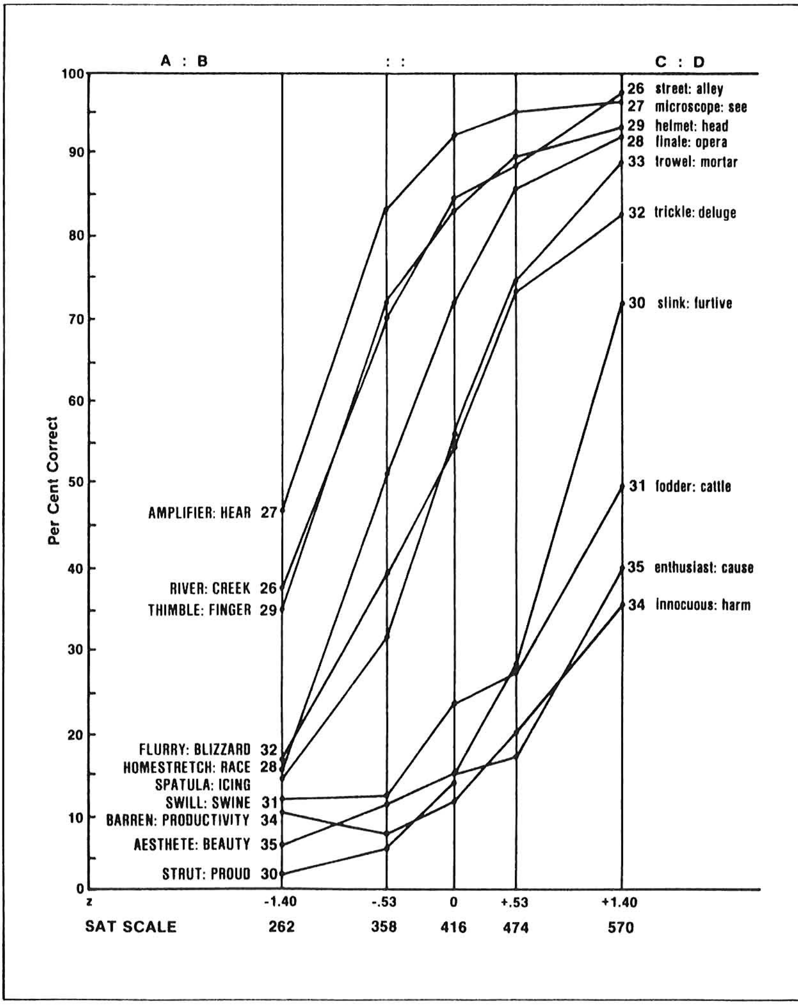

FIG. 8.6. Item characteristic curves for 10 verbal analogies items from a form of the SAT-V (from Carroll, 1980).

arises from the complexity of the rules that are the bases of the analogies. Consider, for example, the hardest of these items:

34. BARREN:PRODUCTIVITY:: (A) torrid:warmth
 (B) innocuous:harm (C) aberrant:change
 (D) prodigal:reform (E) random:originality

The words in the lead, BARREN and PRODUCTIVITY, are not particularly difficult words. The rule relating them is fairly complex: BARREN is an adjective, and PRODUCTIVITY is an abstract noun that signifies a property opposite to that of BARREN. The examinee has to find a choice that correctly exemplifies this relation. From the item analysis data we find that alternative C, *aberrant: change,* is rather tempting, as is also alternative D, *prodigal: reform,* and without careful thought they might appear to exemplify the rule. Only alternative B, *innocuous: harm* correctly exemplifies the rule, but it does so in a fairly subtle way. Unfortunately it would be difficult, though perhaps not impossible, to establish a metric for the difficulty of rules used in verbal analogies items. I would think, however, that it might be possible to make the construction of verbal analogies tests more of a science and less of an art by devoting deliberate attention to constructing items according to a metric for rule-difficulty.

COGNITIVE PSYCHOLOGY AND TASK DIFFICULTY

Much of the current research in cognitive psychology is devoted essentially to finding what elements or aspects of cognitive tasks make them easy or difficult, and in this way the work is directly relevant to test construction and interpretation. One can find many examples, and I can mention only a few.

There has been considerable investigation concerning attributes of tasks used in tests of spatial abilities. Pellegrino and Kail (1982), for example, consider tasks used in tests of two fairly distinct spatial aptitudes—Spatial Relations and Spatial Visualization. In the case of Spatial Relations, the task attributes that chiefly make for difficulty (either in speed or accuracy of response) are angular disparity and familiarity of stimuli. Pellegrino and Kail (1982) conclude on the basis of developmental studies that "individual differences in spatial aptitude are initially associated with basic encoding and comparison processes, that such differences persist over development, and that the differences are then accompanied by additional differences in the speed of mental rotating or transforming the information that has been encoded" (p. 333). In the case of Spatial Visualization, some of the task attributes that make for difficulty are rotation, displacement of elements, and number of stimulus elements. Considering these facts, these authors conclude that "skill in a visualization task such as the form board is related to the speed and quality of the stimulus representation that is achieved" (p. 354).

Another example is the work of Goldman and Pellegrino (1984) on inductive reasoning tasks. They find, among other things, that "the visual or semantic complexity of a particular item affects the degree to which general system characteristics such as working memory and executive monitoring strategies become important cognitive components of performance" (p. 193).

Similarly, I would interpret the work of Rips (1984) on deductive reasoning as an attempt to identify what elements in certain reasoning tasks—verification of arguments containing the connectives *and, or, if . . . then,* and *not*—cause difficulties for subjects. Rips found stable differences between subjects in their handling of rules of reasoning. Although Rips does not present data allowing this direct interpretation, I would speculate that his data suggest that deductive ability can be defined in terms of knowledge of and ability to use an increasingly more complicated set of deductive rules.

APPLICATION OF THE THEORY TO COGNITIVE ABILITY FACTORS

Over the past several years, I have devoted my attention to surveying and in many cases reanalyzing data from the factor-analytic literature in an attempt to determine what the major dimensions of cognitive ability are. I am aware of many of the limitations of factor analysis—they have been pointed out many times. Nevertheless, I have been pursuing my survey on the conviction that if adequate correlational data are uniformly subjected to presently acceptable methods of factor analysis, the results will be more meaningful, consistent, and interpretable than they have appeared to be in the past. I am now approaching the final stages of my survey, and while I am not ready to offer definite conclusions, I now perceive a "light at the end of the tunnel" that appears to confirm my convictions.

One conclusion that now seems evident, however, is somewhat contrary to my initial expectations. My original expectation was that I could identify, from the literature, a fairly large number of factors of ability—not as many as Guilford (1967; Guilford & Hoepfner, 1971) had postulated and claimed to demonstrate—but at least a few more than French (1951) and Ekstrom (1979) had listed in their reviews of the factor-analytic literature. My present view is that there are not more than about thirty distinct, identifiable factors of cognitive ability, and of these, many are of a fairly specific nature and of little importance. The factors that appear over and over in my reanalyses are mostly those originally identified by Thurstone (1938) and other early investigators. Among the first-order "primary" factors that I believe can be confidently identified are Thurstone's Induction, Deductive Reasoning, Verbal Comprehension, Spatial Relations, Visualization, Closure, Perceptual Speed, Associative Memory, Word Fluency, and Memory Span. In addition, there is fairly solid evidence for a series of "second order" broad factors, as identified by Cattell, Horn, and others (e.g., Hakstian & Cattell, 1978): factors of "fluid intelligence," "crystallized intelligence," "general visual perception," "general auditory perception," "general speed," "general memory capacity," and "general idea production." Even some of these factors tend to be correlated, a fact that suggests that Spearman (1927) was correct in asserting the existence and importance of a "general" factor, "g". My analytic procedures assume a hierarchical model such that some factors are of greater generality and applicability than others. The hierarchical model usually results in specifying two or more independent sources of significant variance for a given variable, that is, variance from a primary factor and also variance from a second- or higher-order factor.

Earlier, I suggested that the person characteristic model as illustrated with data from the pitch discrimination test and the block counting test could be transferred or applied to *any* ability. My factorial results, however, pose certain problems for this suggestion.

First, not all factors appear to be characterizable in terms of tasks of varying difficulties. Many, for example, refer mainly to the *speed* of performance of simple cognitive tasks, like, for example, the comparison of stimuli, as in the Perceptual Speed factor. It is not immediately clear how the person characteristic function model can be applied to such factors, unless certain modifications are made in the model. One way of doing this is to utilize individual variation in speed of response over trials as a basis for developing the person characteristic function. A person of a given degree of ability would have an average speed, but the probability of exceeding a given rate would decrease as the baseline value increases. The general idea is illustrated in Fig. 8.7.

The other problem posed by factorial results is the fact that most variables show multiple sources of variance—at least two, as I have mentioned. On the average, I find that about half the common variance of a variable comes from a primary or first-order factor, and the remainder from higher-order factors. This means that many tasks can be supposed to have at least two sources of difficulty—one from a primary factor and one from a higher-order factor, such as a general factor. It would be interesting to work out the implications of this fact for the person characteristic function. I suspect that it means that PCF curves will be somewhat attenuated, i.e., with flatter slopes, when tasks have multiple sources of difficulty. Nevertheless, it may still be possible to separate these effects.

For example, suppose we are concerned, as we should be, with the source of difficulty due to a general factor. That is, independent of the effects of a particular primary factor, what makes a task difficult if it also has a high loading on a

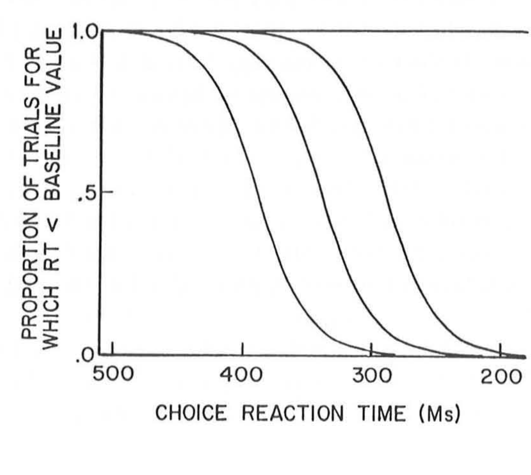

FIG. 8.7. Hypothetical person characteristic curves for three individuals (slow, average, fast) on a speed ability (e.g., choice reaction time).

general factor? If we could find this out, it would contribute to the interpretation of the nature of the general factor. One possibility is deliberately to select tasks that have loadings on different independent primary factors, and then study the person characteristic function for such tasks and the task attributes that function to make them load on a general factor. I am not aware that such an idea has ever been tried. I intend at least to work out the theoretical model by which this might be accomplished, or to determine whether or not it might be accomplished at all.

DISCUSSION

The major points I have been trying to emphasize are the following:

1. The existence of an ability can be demonstrated when it can be shown that for any individual, there is a systematic, monotonic, and close relation between the individual's probability of correct or satisfactory performance and the difficulties of a series of tasks, and when there are variations over individuals in the parameters of this relation.
2. The ability is defined in terms of the attribute or attributes of the tasks that give rise to differences in task difficulty.
3. This formulation, or one closely similar to it, is applicable to any cognitive ability.
4. Cognitive psychology can be of help in the definition of cognitive abilities by investigating what attributes of tasks make for differences in the accuracy or speed with which individuals can perform those tasks, because such attributes are involved in the definition of abilities. Further, knowledge of task attributes can lead to inferences about the psychological processes that are called for in performances, and thus about the psychological processes that underlie a given ability.

A corollary of this formulation is that effects of education, training, or other forms of intervention can be indexed by changes in the position parameter of an individual's person characteristic function. A significantly positive effect of learning or an educational intervention, for example, would be exhibited in a significant increase in the individual's threshold of performance along the task difficulty scale.

The person characteristic function (PCF) model can be shown to apply at least in a number of "simple cases." Probably it could be shown to apply to most of the major types of ability that have been identified. Undoubtedly certain complications would arise in more complex cases. Among these complications are:

1. The possibility that task performance may be a function of more than one ability.

2. The possibility that tasks could be performed through different ''strategies'' or approaches.

3. The possibility that at least some abilities, especially those representing educational achievements, involve results of specific learning. The fact that individuals may vary in what particular learnings they may have achieved, independent of the overall difficulty of those learnings, may present problems in applying the PCF model to certain kinds of abilities.

No doubt it would be fruitful to study the problems posed by these complications, but I believe that such studies would be appropriate only after considerable success has been achieved in applying the PCF model to ''simple cases.'' Since this has been done thus far to only a limited extent, there is a wide field of problems open for examination.

One final remark: I have only intimated how all this might help in better test construction. I will try to be more explicit: We can make better tests of abilities by paying more attention to the task characteristics that make for item ease or difficulty and to the role of such task characteristics in defining the abilities we seek to measure.

REFERENCES

Carroll, J. B. (1980). Measurement of abilities constructs. In U. S. Office of Personnel Management and Educational Testing Service, *Construct validity in psychological measurement: Proceedings of a Colloquium on Theory and Application in Education and Employment* (pp. 23–41). Princeton, NJ: Educational Testing Service.

Carroll, J. B. (1983). The difficulty of a test and its factor composition revisited. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A Festschrift in honor of Frederic M. Lord* (pp. 257–283). Hillsdale, NJ: Lawrence Erlbaum Associates.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American Heritage word frequency book.* Boston: Houghton Mifflin.

Ekstrom, R. B. (1979). Review of cognitive factors. *Multivariate Behavioral Research Monographs,* No. 79–2, 7–56.

French, J. W. (1951). The description of aptitude and achievement tests in terms of rotated factors. *Psychometric Monographs,* No. 5.

Goldman, S. R., & Pellegrino, J. W. (1984). Deductions about induction: Analyses of developmental and individual differences. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, Vol. 2* (pp. 149–197). Hillsdale, NJ: Lawrence Erlbaum Associates.

Guilford, J. P. (1941). The difficulty of a test and its factor composition. *Psychometrika, 6,* 67–77.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Guilford, J. P., & Hoepfner, R. (1971). *The analysis of intelligence.* New York: McGraw-Hill.

Hakstian, A. R., & Cattell, R. B. (1978). Higher-stratum ability structures on a basis of twenty primary abilities. *Journal of Educational Psychology, 70,* 657–669.

Johnson, E. S., & Meade, A. (1982). Personal communication.

Lohman, D. F. (1979). *Spatial ability: Individual differences in speed and level* (Tech. Rep. No. 9). Stanford, CA: Aptitude Research Project, School of Education, Stanford University.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mosier, C. I. (1941). Psychophysics and mental test theory, II. The constant process. *Psychological Review, 48,* 235–249.

Pellegrino, J. W., & Kail, R., Jr. (1982). Process analyses of spatial aptitude. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence,* Vol. 1 (pp. 311–365). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rips, L. J. (1984). Reasoning as a central intellective activity. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence,* Vol. 2 (pp. 105–147). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shepard, R. N., & Feng, C. (1972). A chronometric study of mental paper folding. *Cognitive Psychology, 3,* 228–243.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: Macmillan.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs,* No. 1.

Walker, D. A. (1931). Answer pattern and score scatter in tests and examinations. *British Journal of Psychology, 22,* 73–86.

Walker, D. A. (1936). Answer pattern and score scatter in tests and examinations. *British Journal of Psychology, 26,* 301–308.

Walker, D. A. (1940). Answer pattern and score scatter in tests and examinations. *British Journal of Psychology, 30,* 248–260.

Wilson, D., Wood, R., & Gibbons, R. D. (1984). TESTFACT [computer program]. Mooresville, IN: Scientific Software.